IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 1, 2021, accepted November 4, 2021, date of publication November 8, 2021, date of current version November 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3126486

Towards the Mobile Detection of Cervical Lesions: A Region-Based Approach for the Analysis of Microscopic Images

ANA FILIPA SAMPAIO, LUÍS ROSADO, AND MARIA JOÃO M. VASCONCELOS
Fraunhofer Portugal AICOS, 4200-135 Porto, Portugal

Corresponding author: Ana Filipa Sampaio (ana.sampaio@fraunhofer.pt)

This work was supported by project CLARE: Computer-Aided Cervical Cancer Screening, with operation code POCI-01-0145-FEDER-028857, cofounded by European Regional Development Fund (FEDER) through Operational Programme for Competitiveness and Internationalisation (COMPETE 2020) and National Funds through the Foundation for Science and Technology (FCT/MCTES).

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Hospital Prof. Doutor Fernando Fonseca (Approval ID: 16/2019).

ABSTRACT Given the current prevalence and impact of cervical cancer worldwide, many technological developments focused on automating the screening process have arisen recently. Nonetheless, there is still a clear need for affordable, portable and automated IoT-based solutions to expand the coverage of current cervical screening programs worldwide. This is particularly relevant for lower-resource countries, which account for 88% of all cervical cancer-related deaths. This work proposes a low-cost, smartphone-based microscopy device for the analysis of liquid-based cytology samples, through autonomous image acquisition and automated identification of cervical lesions. Different deep learning models for object detection were separately optimised and compared to select the most adequate network architecture. Transfer learning from a similar application domain - conventional cytology - was also investigated as a way of improving the robustness of the analysis pipeline, as well as overcoming the limitations of the mobile-acquired image dataset specifically collected and manually annotated by specialists under the scope of this work. In this process, a detection performance benchmark in the SIPAKMED dataset - test mean average precision (mAP) of 0.37798 and average recall (AR) of 0.63651 - was reported for the first time. Although further improvements are required for its integration in a computer-aided diagnosis system sufficiently reliable for deployment in a clinical context, the explored approach exhibits promising results (cross-validation mAP of 0.20315, AR of 0.46572 and analysis time of 4 minutes per cytological sample), corresponding to a step forward in the development of a cost-effective mobile IoT framework that supports cervical lesion screening.

INDEX TERMS Artificial intelligence, computer aided diagnosis, deep learning, Internet of Things, knowledge transfer, microscopy, object detection, telemedicine.

I. INTRODUCTION

Ranking as the fourth most common cause of cancer incidence and mortality in women worldwide, cervical cancer continues to constitute a major public health problem. In 2018, approximately 84% of all cervical cancers and 88% of all deaths caused by cervical cancer occurred in lower-resource countries [1]. Moreover, the mean age at diagnosis of cervical cancer is quite low compared with that of most other major cancer types, generating a proportionally greater loss of life-years.

In order to reduce mortality rates, the worldwide adoption of both early detection and screening programs is essential [2]. Examples of screening methods recommended by the World Health Organization (WHO) are: i) visual inspection of with acetic acid (VIA); ii) cervical cytology through conventional (PAP) test or liquid-based cytology (LBC); iii) Human papillomavirus (HPV) testing for high-risk HPV types. The first method has been more adopted in low-resource settings, due to its cost, in spite of its limited accuracy for

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

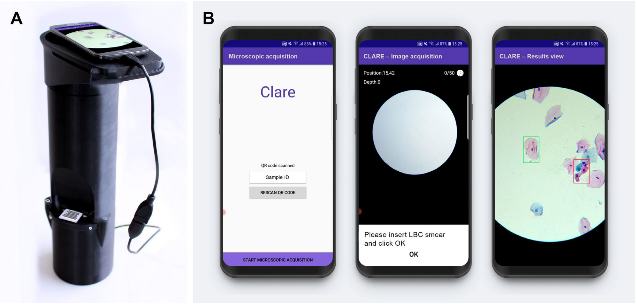
the detection of pre-cancerous lesions. The second method has been the standard method for screening, being linked to drastic reductions of mortality rates after its adoption in many countries, with LBC being used in more developed countries and conventional Pap tests otherwise. Finally, in more recent years HPV tests have been used for screening, either alone or in combination with Pap tests, since most cervical cancers are caused by HPV.

Over the last years, efforts have been made to support the screening actions in order to facilitate the diagnosis and coverage of screening programs, namely through the development of automated microscopy solutions such as the ThinPrep Imaging System (TIS) and the BD FocalPoint GS Imaging System (FocalPoint) [3]. However, their high prices are prohibitive for widespread usage and the necessity of having an affordable solution to automatically acquire images from cytology samples remains, as well as the development of computer-aided diagnosis systems to assist in the identification of cervical lesions [4]. As a result, the search for new and innovative solutions that are simultaneously effective and affordable has the potential to improve cervical cancer screening, in particular: (i) increase the sensitivity and specificity of manual screening; (ii) decrease the cytotechnologists' workload; (iii) reduce the cost of screening programs; and (iv) reduce the disease incidence and mortality rate [4].

Regarding affordable and automated alternatives to conventional microscopes, Fraunhofer AICOS has been working in the development of a fully automated 3D-printed smartphone microscope, named μSmartScope [5]. This prototype is an affordable and automated alternative to conventional microscopes, tailored to effectively support microscopy-based diagnosis in areas with limited access to healthcare services. Through the usage of a motorised stage entirely powered and controlled by a smartphone, the μSmartScope allows autonomous acquisition of microscopic images, with the ultimate goal of decreasing the burden of manual microscopy examination. The μSmartScope also aims at reducing dependence on experts in microscopy diagnosis available on-site by allowing straightforward integration with Artificial Intelligence (AI). The device was already used for automated analysis of malaria-infected blood smears [6], [7] and is currently being redesigned for cervical cancer screening, to achieve a solution that fulfils the requirements for the accurate microscopic examination of liquid-based cervical cytology samples (Figure 1).

The present work addresses the existing open problems in cervical cancer screening through the proposal of a framework based on the Internet of Things (IoT) technology that integrates the μSmartScope device for the acquisition of images of microscopic fields of view from LBC samples with deep learning models for the automated detection and classification of cervical lesions in the mobile-acquired microscopic fields.

Despite the growing processing capacity of smartphones and the subsequent possibility of mobile execution of the developed algorithms, the computational complexity of the state of the art detection models hinders their local deployment. Hence, the proposed system should integrate the mobile image acquisition with a cloud-based implementation of the processing pipeline, highlighting the importance of the IoT for the successful integration of the AI algorithms in a practical decision support system.



Please insert LBC smear and click OK

FIGURE 1. Mobile-based framework for the detection of cervical lesions: (A) μSmartScope with smartphone attached and LBC sample inserted; (B) smartphone application screenshots exemplifying the usage of the restructured solution (from left to right): (i) cervical sample insertion; (ii) optic disk alignment and start of the image acquisition through the smartphone app; and (iii) example of possible visual feedback of the automated lesion detection.

For the development of the cervical lesion identification models, three object detection architectures based on deep neural networks were optimised and compared, aiming to identify the best model for this application domain. Due to the limitations of the dataset collected with the portable microscopy equipment, a more thorough study was performed on a public dataset of conventional cytology (the SIPAKMED dataset [8]) beforehand, to identify the most promising models and obtain a performance baseline for the detection pipeline. The potential of knowledge transfer between the conventional cytology domain and the liquid-based application was assessed as a means to improve the robustness of the detection models, and the reliability of the final model was inspected in detail.

Besides describing an innovative mobile-based IoT framework capable of assisting the clinicians' cervical lesion diagnosis process, this work comprises other developments that contribute to the progress of automated cervical cancer screening research. The study conducted in the SIPAKMED data provides a performance benchmark on the cell detection task (yet to be reported in the literature), also introducing transfer learning from a close application domain as a way of overcoming private data limitations. In addition, the construction of a novel dataset of LBC microscopic fields acquired in mobile settings, in association with the efforts undertaken to tackle its shortcomings and the recommendations put forward based on the challenges encountered, provide an important foundation for future research in this field.

II. RELATED WORK

One of the main tasks encompassed by the analysis of cervical samples is the **identification of cervical lesions in the microscopic fields**. As a consequence, this is a pivotal step for the development of a successful computer-aided cervical cancer screening system. In spite of the myriad algorithmic methodologies to achieve it reported in the literature [9], only the main lines of research are mentioned in this section.

Many works attempt to fulfil this endeavour through the **segmentation of the cells** present in the image, with the goal of further categorising them according to their abnormality. Although the main focus of this step is to locate the cells that should be analysed in detail, it can also be the basis to compute cell features with a relevant clinical meaning, such as the shape and dimensions of cells and their respective inner structures.

Some authors investigated the application of traditional image analysis techniques for this purpose. An example is the work of Byju *et al.* [10], which resorts to a customised Laplacian of Gaussian (LoG) filter to detect the contours of the cell's nuclei. More complex strategies are presented in [11], proposed in the context of an international challenge, and seek to segment individual cells, with a special emphasis on separating the individual cells that may appear aggregated and partially overlapping in clumps. The best-performing pipelines include the segmentation of the nuclei structures as a way of establishing inner shape priors of the cells, used as a basis for refined active contour algorithms, responsible for the segmentation of the cytoplasm regions. In some of the methodologies described, the nuclei delineation step takes advantage of the darker staining of these regions, applying iterative, progressively increasing, intensity thresholds that take into account other relevant properties of the segmented objects (e. g. their area and eccentricity) in this process.

Alternative strategies for cell segmentation rely on machine learning models, applied to determine if each image pixel belongs to the cell region or the background. In this line of research, the works of [12] and [13] employ pixel-wise classifiers to distinguish nuclei, cytoplasm and background pixels with reasonable success; still, these methods were only applied to single-cell images and not to broader microscopic fields, limiting their standalone usability.

Other studies were reported for the inspection of individual cell instances, mostly for their **categorisation in abnormality or lesion levels**. Many of these pipelines make use of the Herlev dataset [14], comprised of images of cervical cells differentiated in normal and abnormal instances, and stratified according to the abnormality levels contemplated by the prior classification scheme for cervical intraepithelial neoplasia (CIN) [15]. Most of the strategies reported focus on the extraction of relevant features from the cell images, using conventional machine learning models to predict their CIN level. In [16], statistical features are combined with the cell's perimeter and a rotation-invariant feature - the global significant value (GSV), representative of the overall intensity variation of the cell's image - to generate the feature vector used to feed the classifier. Four classification models- k-NN (k- nearest neighbours), Bayesian network, J48 tree and multi-layer perceptron (MLP) - are compared, with k-NN yielding a performance superior to the most state of the art approaches (88.45%), while keeping a lower computational time (average of 752.05 ms). Due to their relevance in the clinical reasoning process, nuclei features are also considered by some methodologies, such as [17], which assessed the impact of using morphometric (geometrical and colour) and texture features from the nuclei regions. The authors demonstrated that the combination of both types of features enhanced the ability of the model to distinguish a subset of the Herlev dataset's classes. In [18], features from both the nuclei regions and the whole cells are used to separate the cells according to all the Herlev classes, even though some of them are merged, in light of the similar nuclei features exhibited. To tackle the multiplicity of features computed - which might excessively increase the computational requirements of the classification algorithm -, these two methodologies apply principal component analysis (PCA) as a way of reducing the dimensionality of the feature vector. In both cases, several classification models are compared, with support vector machines (SVM) yielding the most accurate performances.

Other strategies propose the usage of convolutional neural networks (CNNs) for feature extraction instead of computing hand-crafted features, as an attempt to automate this step of the pipeline. In [19], image features are extracted using a CNN trained in the Herlev cell instances and three options of classifiers - a softmax layer, an SVM and a gentle ensemble of decision trees (GEDT) - are assessed to identify the cervical lesion level, with GEDT achieving the most encouraging results. The work of [20] merged the Herlev data with a private dataset to develop a model for the classification of cell images, obtained by modifying the structure of a Resnet network to make it more adequate for resource-constrained settings; the created lightweight model exhibits a competitive performance in relation to most state of the art methods for this task. Kwon *et al.* [21] attempt to bridge the gap between the CIN convention and the Bethesda System (TBS, currently employed in clinical settings) [22] through the development of a tailored CNN model to classify cell regions as normal, low-level lesions or high-level lesions. For that, they assembled a novel dataset characterised by the inclusion of both single cells and clumps. Their results show the potential of CNNs for the analysis of cervical images.

While the classification of single cell instances is a critical step for the identification of cervical lesions and the Herlev dataset enabled the development of extensive research for this task, it is not able to provide a direct analysis of whole microscopic fields of view from cytology samples. In consequence of this and of the lack of a benchmark dataset for the identification of abnormal cells in microscopic fields, some works rely on private datasets to develop more complete pipelines. Several of the most recent approaches make the most of deep learning (DL) algorithms, employing **CNN-based object detection models** that are able to locate regions of interest and provide their corresponding classification label through a unified architecture. In [23], a Faster R-CNN model [24] is implemented to detect the nuclei of cervical cells and separate them according to five classes, including normal and abnormal epithelial cells, as well as other types referring to inflammatory cells and debris. Three backbone CNNs are tested for the detection model - VGG16, Resnet50 and Resnet101 - with Resnet50 yielding the best performance overall, in spite of its long inference times. The Faster R-CNN architecture is also the basis for the framework proposed in [25], which uses microscopic images acquired at two magnification levels to train a model capable of detecting and classifying cervical lesions from a subset of 3 (out of 5) abnormal classes from the TBS system. The model was evaluated not only in terms of object-level performance, but also concerning its overall diagnosis outcome per clinical case, determined as the most severe lesion level found in all the fields of each case. Even though the model achieved promising results overall, it exhibits a moderate specificity and cannot detect the remaining types of lesions from the TBS, not included due to the insufficiency of data available. Liang *et al.* [26] address the data limitations through the usage of a comparison-based classifier for each object. This model resorts to a Faster R-CNN model to identify and classify cervical lesion instances, but generates the final classification result by matching the convolutional features extracted by the detection model with the ones obtained for prototype images of each class; the label of the most similar prototype class is considered as the final one. Despite the performance gains produced by the introduction of the comparison-based classifier, these were not sufficient to overcome the critical limitations of the private dataset used, hindering the achievement of results worth highlighting.

Some detection-oriented frameworks focus on the improvement of the overall diagnosis performance of the developed systems. The work of [27], based on a YOLO v3 detection model, proposed the consideration of intermediate classes for each lesion level and the usage of several post-detection classifiers that take advantage of the properties of the surrounding regions and of the cells' nuclei to reduce the number of false detections and identify other infectious diseases that might be present. This framework, developed using a massive multi-centre dataset, corresponds to one of the most complete systems reported in the literature for cervical lesion diagnosis.

Notwithstanding the optimistic results attained by frameworks based on the Faster R-CNN meta-architecture, the search for the most adequate detection model is still an open problem, highly dependent on the type of data available. Moreover, given the complexity of DL-based detection models, many approaches make use of transfer learning methodologies to accelerate the learning process, exploiting models pre-trained on public datasets from other image analysis tasks (e.g. the ImageNet [28], PASCAL Visual Object Classes [29] and Common Objects in Context [30] - COCO - databases). Still, none of these approaches investigated the possibility of transferring knowledge of a closer application domain. Both these shortcomings motivated the development of the proposed system, focused on the identification of the most suitable model for a public cervical cytology database - the SIPAKMED dataset [8] - and on the usage of the models trained on this data to address the limitations of a private dataset of images of cervical cancer lesions acquired using a distinct type of microscopic preparation.

III. DATASETS AND DATA PREPARATION

This section describes the used image datasets and the respective data preparation procedures. The creation of the mobile HFF regions dataset, detailed in section III-A, was motivated by the lack of a benchmark dataset for the localisation of cervical lesions in microscopic fields of liquid-based cytology samples and their stratification according to the abnormality levels of the Bethesda system. Although there are many public datasets based on images of cervical cytology, some of them only provide images of previously separated single cells and classify them according to CIN levels instead of the Bethesda system's classes (e.g., the Herlev dataset [14]), while the ones that include images of microscopic fields of view either do not provide abnormality classification labels for each cell/region [31] or separate the existing cells in relation to their type and not abnormality level (as it is the case of the SIPAKMED dataset, characterised in section III-B), besides including images from conventional cytology preparations instead of LBC samples.

A. MOBILE HFF REGIONS DATASET

The mobile HFF regions dataset is a mobile-acquired image database specifically created under the scope of this work. It is comprised of 21 LBC samples from different clinical cases, supplied by Hospital Fernando Fonseca. The images were acquired using the μSmartScope prototype coupled to a smartphone, being manually annotated by an experienced specialist in terms of abnormal cells or cell aggregates indicative of cervical lesions. The annotations are provided in the form of bounding boxes enclosing the abnormal regions, along with a classification label indicating each region's lesion level according to the Bethesda System's convention [22]: atypical squamous cell of undetermined significance (ASC-US); low-grade squamous intraepithelial lesion (LSIL); atypical squamous cell, cannot exclude high-grade lesion (ASC-H); high-grade squamous intraepithelial lesion (HSIL); and squamous cell carcinoma (SCC). Figure 2 provides examples of the several classes of regions included in the dataset (vertically organised), showing the variability of structures that may be associated with the same lesion level and the similarity between some cells of consecutive lesion levels.

Each sample of the mobile HFF regions dataset comprises 79 ± 21 ($mean \pm std$) images, with only a part of those images actually containing annotations of abnormal regions. Table 1 presents the sample distribution, as well as the number of annotations per lesion class. In this dataset, there are also two cases with a normal diagnosis outcome and three inadequate samples. Furthermore, as a result of the progressive nature of the disease, each sample may contain annotations with abnormal regions of multiple lesion levels.

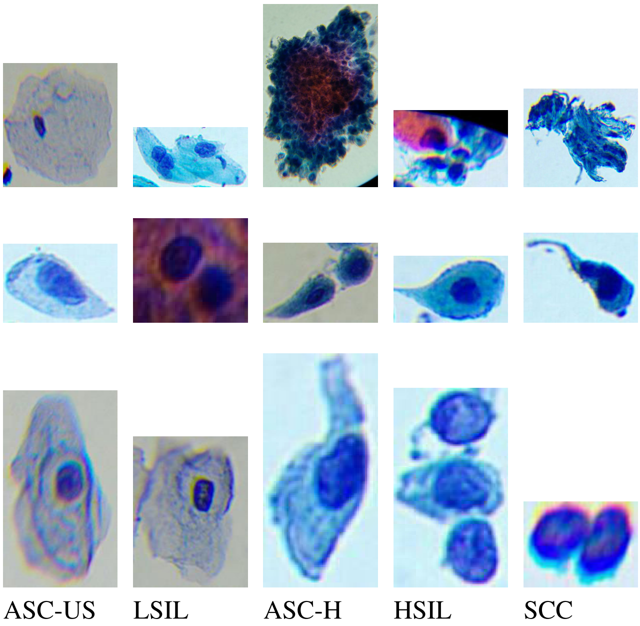


FIGURE 2. Examples of the five region classes considered in the mobile HFF regions dataset.

TABLE 1. Mobile HFF regions dataset sample and annotation distribution (training, test and total).

Number	ASC-US	LSIL	ASC-H	HSIL	SCC	Total
Samples	4	3	4	3	2	16
Train annot.	352	58	79	203	13	705
Test annot.	125	38	30	29	0	222
Total annot.	477	96	109	232	13	927

1) SUBSET DIVISION

The separation of the data instances in the train, validation and test subsets was performed in 2 phases:

i) An initial overall train/test split, performed at the sample level, in which all the images associated with a given source sample (i. e., from the same patient) were kept in the same subset. The sample-wise separation ensures that the test data does not contain repeated fields of view or images very similar to the ones used to train the algorithm. Samples used for training and test purposes were separated manually, based on the visual observation of the types of structures of each class existing in the different samples. This process was also performed in accordance with the per-class distribution of the dataset and aiming at an 80/20 train/test proportion. The limited amount of annotated images, allied to the multiplicity of morphological properties exhibited by the annotated abnormal regions (even for the same lesion level), motivated this choice of division instead of a random stratified split, with the purpose of ensuring that the dataset's diversity is similarly represented in the training and test subsets.

ii) The creation of several train/validation subset splits by means of a stratified k-fold cross-validation procedure applied separately to each training sample. The partition of the train and validation images was performed separately for each sample to ensure that a similar number of images was included for all the overall diagnosis classes, to preserve the dataset's class distribution. The images of each sample were divided in k mutually exclusive subsets, and each split was obtained by selecting one of the subsets for validation and the remaining ones for training; a k value of 5 was selected, as it allowed a reasonable amount of validation instances while providing sufficient training images. The final training and validation sets of a given split resulted from the combination of the subsets of all the samples divided for that split.

2) IMAGE PATCHES EXTRACTION

To obtain images of fixed dimensions, required for the application of some of the detection models and to restrain the computational resources used during training, the images acquired with the μSmartScope were divided into adjacent patches. The extraction of each patch was executed taking into account the annotated regions contained in its area, through a procedure detailed in Appendix A and illustrated in Figure 3. These steps were applied after pre-processing the acquired images to segment the optic disk (according to the steps described in [7]) and crop the main field of interest in accordance with the segmented region. The dimensions of the extracted patch images were one of the settings optimised during the tuning process, as later specified in Section IV-B.

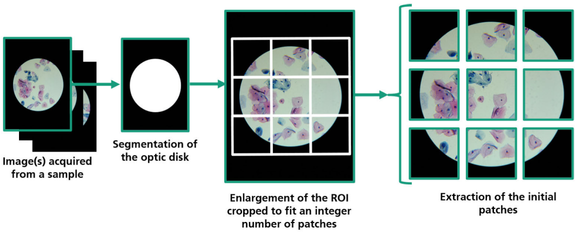

Image(s) acquired from a sample
Segmentation of the optic disk
Enlargement of the ROI cropped to fit an integer number of patches
Extraction of the initial patches

FIGURE 3. Patch extraction process.

3) ADDRESSING THE MAIN LIMITATIONS OF THE DATASET

Although the mobile HFF regions dataset is fairly even concerning the proportion of samples for each diagnosis outcome, the distribution of abnormal regions is not balanced for all lesion levels, as it is clear from Table 1. Furthermore, the amount of clinical cases of each class is scarce, and the total number of microscopic fields with abnormal regions may be insufficient to train complex neural network models.

In addition to this, the number of patch images without annotated objects (henceforth referred to as empty patches) surpasses substantially the number of patch samples that contain actual regions of interest: the annotated patches comprise approximately $3 - 7\%$ depending on the size of the extracted patches. This imbalance could lead to many training steps being performed using mainly images without bounding boxes from which the network can learn, dampening the learning process.

Therefore, to address these shortcomings, efforts were made in terms of pre-processing operations, as described next:

i) **Merging similar classes:** Taking into consideration the substantial under-representation of the SCC class, and bearing in mind that the clinical diagnosis flow is similar for both SCC and HSIL types of lesions, they were merged in a single class, designated as HSIL-SCC.

ii) **Down-sampling empty images:** To avoid the predominance of training instances with no object information and ensure that the dataset comprised an even proportion of empty and annotated fields, empty patches of each clinical sample were down-sampled through the phases outlined in Appendix B.
With this process, the percentage of annotated patches increased from $3 - 5\%$ to approximately $60 - 67\%$ (depending on the patch size), yielding a slightly superior proportion of annotated instances, which can be advantageous to the training process.

iii) **Getting a more uniform class distribution:** To compensate for the class imbalance that critically affects the ability of the model to learn proper representations of the under-represented classes, the number of training instances of those classes was increased through data augmentation transformations. Several image transformations were applied: geometrical transformations, namely vertical and horizontal flips, as well as rotations of 90°, 180° and 270°; intensity transformations, in particular blur and sharpening, to mimic the varied focus levels of the structures naturally found in the images, and gamma correction applied to each RGB channel separately (which accounted for three distinct transformations).
The calculation of the number of additional images needed for each class was performed considering the difference between the representation of the dominant class and the representation of the adjusted class. To simplify the computations, this analysis was conducted according to the number of patches with objects of each class, assuming an average of a single object per annotated image (which was verified through an inspection of the patch annotations).

B. SIPAKMED DATASET

The SIPAKMED dataset [8] is composed of 966 images of conventional cytology samples acquired using a CCD camera (Infinity 1 Lumenera) adapted to an optical microscope (OLYMPUS BX53F), as well as expert annotations concerning the cytoplasmatic and nuclear contours of each cell type. The images contain 5 types of cells - superficial/intermediate (Sup.-Int.), parabasal (Parab.), koilocytotic (Koiloc.), dyskeratotic (Dysk.) and metaplastic (Metap.), corresponding to distinct types of epithelium and including abnormal and normal cell classes. Despite the diversity of cell types that are a part of the dataset, each image only encompasses annotations of a specific cell class, even if the captured microscopic field includes cells of some of the other classes. Some examples of each class are included in Figure 4, and the per-class distributions of the images and cell objects are presented in Table 2.

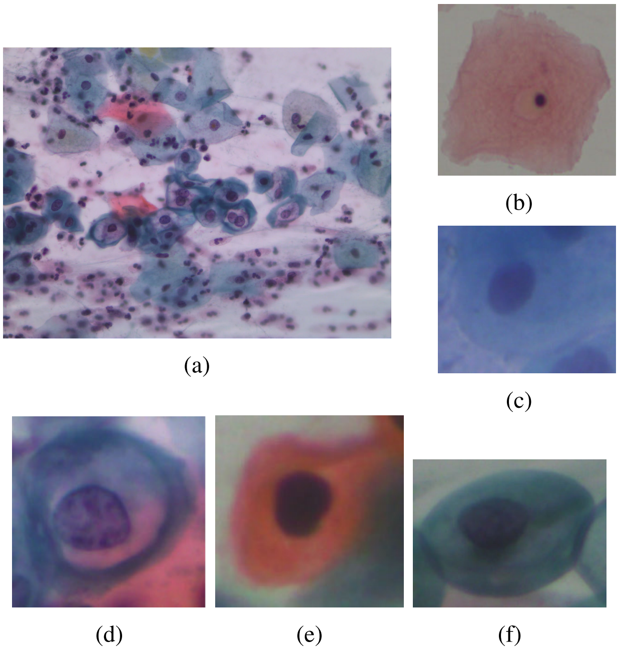


FIGURE 4. Images from the SIPAKMED dataset. Figure (a) is an image with koilocytotic cells. The remaining images are examples of the five cell types considered: (b) superficial/intermediate; (c) parabasal; (d) koilocytotic; (e) dyskeratotic; and (f) metaplastic.

TABLE 2. SIPAKMED dataset image and cells distribution.

Number	Sup.-Int.	Parab.	Koiloc.	Dysk.	Metap.	Total
Images	126	108	238	271	223	966
Cells	813	787	825	793	813	4049

There is a slight imbalance in the number of images of some classes, in particular of parabasal and superficial-intermediate cells; yet, each of these images contains more cell instances than the images of the remaining classes, resulting in a final well-balanced object distribution and preventing the need of using data augmentation strategies for balancing purposes.

Although the provided annotations correspond to the contours of the cells and nuclei, the detection architectures require ground truth annotations as bounding boxes. As such, the ground truth bounding boxes to develop the detection algorithms were obtained as the smallest rectangular boxes that encompassed the whole cytoplasmatic contour of each cell. To ensure that the pipelines were similar for both datasets, the SIPAKMED images were subjected to most of the pre-processing steps described in Section III-A2 for the mobile HFF regions data. Yet, on account of the different properties of the conventional cytology images, some operations were not applied, namely the segmentation of the optic disk and the augmentation transformations used to harmonise the dataset's distribution.

IV. METHODOLOGY

The main goal of this work was to develop a model able to locate and classify cervical lesions in images of microscopic fields of LBC samples. Despite the pre-processing efforts employed to handle the restricted volume of images and the peculiarities of the mobile HFF regions dataset, these shortcomings still hindered the successful development of robust pipelines based on deep learning models. For this reason, the construction of the main region detection approach was first conducted using the public SIPAKMED dataset [8]. Even though it is composed by conventional cytology samples and its classification labels correspond to cell types instead of abnormality levels, both datasets are from cervical cytology preparations and the corresponding annotations divide them according to morphological properties relevant for the identification of cervical lesions. Thus, two distinct region detection studies were conducted:

1) the search for the optimal model for region detection and identification of the associated type of cells, based on the conventional microscopy samples of the SIPAKMED dataset (using models pre-trained in the COCO dataset);
2) the investigation of the knowledge transfer utility between the two types of cytology preparations, through the application of the best meta-architecture and backbone network resulting from the SIPAKMED studies to the mobile HFF regions images.

A. REGION DETECTION MODELS

In the first study, carried out using the SIPAKMED dataset, several DL models for object detection were explored and compared to identify the most promising approach. The Tensorflow Object Detection API was chosen as a development tool, in view of its vast assortment of pre-trained models available and its straightforward usage. As each detection model is defined by the combination of a detection meta-architecture with a specific backbone network, the selection of the models studied was focused on the inclusion of meta-architecture/backbone combinations that provided varying levels of computational complexity and detection performance, according to the benchmarks reported in the literature [32].

Two backbone networks, used to extract the base convolutional feature maps from the input images, were selected - Resnet50 and Mobilenet v2. Resnet50 [33] is a CNN model that integrates residual connections - shortcuts between feature maps of different depth levels - in some of its 50 layers, with the objective of propagating lower-level image representations to deeper layers of the network, which enables an easier convergence of deeper, more accurate models [34]. On the other hand, Mobilenet v2 [35] is a lightweight network optimised for mobile deployment through the usage of depth-wise and point-wise convolutions, as well as the introduction of a width multiplier setting that allows the adjustment of the network's computational cost without compromising its accuracy. Thus, Mobilenet v2 is characterised by faster inference times, but it might yield slightly worse performances than deeper networks; in contrast, the higher complexity of Resnet50 should lead to more robust detection outcomes.

The selection of the object detection meta-architectures was made taking into account the advantages and features of each algorithm. Faster R-CNN [24] is a two-stage detector with a slower inference speed and a complexity that might be discouraging for mobile settings; however, it presents the most reliable performance in terms of mean average precision (mAP) and detection sensitivity, especially with small objects, so it was chosen to set a performance baseline. SSD [36] is a one-stage detector that presents a promising detection performance in mobile settings, both in terms of mAP and inference speed, but it might exhibit a worse sensitivity to small objects. RetinaNet [37] corresponds to another one-stage detector, trained with a focal loss function that assigns a greater influence to harder instances in the parameter updates, to address the under-representation of some object classes, providing an intermediate level of speed and detection robustness.

Due to the complex structure of the detection models and the high volumes of data required to train them from scratch, this study was based on networks pre-trained in the public Common Objects in Context (COCO) dataset [30], which were fine-tuned to the cervical cytology context. Considering the meta-architecture/backbone combinations with available pre-trained weights in the Tensorflow Object Detection API, three models were singled out for the SIPAKMED detection experiments:

- SSD with a Mobilenet v2 backbone;
- RetinaNet with a Resnet50 backbone;
- Faster R-CNN with a Resnet50 backbone.

As all models are associated with a different meta-architecture, in the remaining sections of this document each model is identified using simply its detection scheme (e. g. SSD with a Mobilenet v2 backbone is referred to as SSD model) for the sake of briefness. To ensure that the optimal settings were found for all models, each meta-architecture/backbone combination was subjected to an appropriate hyperparameter tuning procedure, detailed in further sections. Even though several models were examined for the SIPAKMED studies, to limit the number of necessary experiments, only the best-performing meta-architecture/backbone combination was used for the HFF regions analysis.

B. TRAINING SET-UP

For each meta-architecture/backbone combination, a small set of hyperparameters were optimised. A random search procedure was applied instead of grid search due to time constraints. The hyperparameters used to specify the object proposals generated were established considering the properties of the dataset's regions, while the remaining settings were fixed at the default values used in the pre-training step.

1) RANDOM SEARCH HYPERPARAMETERS OPTIMISATION

Some of the most crucial hyperparameters and training settings - the input image dimensions, the batch size and the learning rate - underwent a random search process, which enables the exploration of a vast range of values while keeping the number of performed experiments low. In this method, the search space is randomly sampled and only a fixed number of combinations of possible values for those hyperparameters are tested and compared [38].

Two input image dimensions were assessed: 320 × 320 and 640 × 640 pixels. These input dimensions were preferred because they were the ones used to train the available pre-trained models that corresponded to patch scales in which the objects of interest display distinguishable scales.

In terms of batch sizes tested (all powers of base 2), only sizes below 32 were taken into account, as the usage of larger batches was not possible due to dynamic memory constraints. Batches with 2 or 4 images were not considered, owing to the insufficient amount of annotated instances with relevant information that would be used in each training iteration in those cases. Thus, there were just two batch sizes compliant with the aforementioned conditions - 8 and 16.

The learning rate values, kept constant for each experiment, were randomly sampled from a logarithmically scaled range between 10^{-6} and 0.01, in order to scan different learning rate magnitudes, as recommended in [39].

Each combination of random values of these three settings specified a separate experiment. On account of the long training times per experiment required, the number of hyperparameter combinations explored was limited to five for each detection model in the SIPAKMED study and three in the case of the HFF regions study. Details regarding the hyperparameter values tested for both studies can be found in Tables 7 and 8 of Appendix D.

2) HYPERPARAMETERS ADJUSTED TO THE DATASET

Other hyperparameters that might be crucial to achieve the proper detection of the regions of interest are the ones associated with the anchor boxes, i.e. the bounding box priors that define the object candidates proposed by the model. These anchors are usually specified in terms of scales and aspect ratios that are combined to generate the box templates applied at each pre-defined image location to extract the possible objects. In the conducted experiments, the anchors' scales and aspect ratios were adjusted to the dataset used, based on the dimensions and shapes of its bounding box annotations. The optimal values for these hyperparameters were determined with a clustering-based approach, through the methodology described in [40]. This strategy was applied independently for each dataset, considering only the data instances assigned for training purposes.

3) OTHER TRAINING SETTINGS AND HYPERPARAMETERS FIXED

The networks were trained for a maximum of 400 thousand iterations, as needed to attain convergence. Despite the large number of training steps, the validation performance of the model was monitored during training, and the model with the best validation performance was regarded as the final model for each experiment. Adam [41] was used as the optimisation algorithm, in virtue of the adaptive learning rate adjustments it provides and its state of the art performance. The loss functions adopted for the localisation and classification tasks were the Huber regression loss [42] and the cross-entropy loss, respectively.

Due to their importance to filter out duplicated predictions for the same object and to their impact on the sensitivity of the algorithm, the post-processing thresholds of the non-maximum suppression stage were fixed at 0.5 for the intersection over union (IOU) and 0 for the object score. A low score threshold was chosen to ensure that most predictions were analysed, given the low confidence scores achieved for many objects.

The remaining hyperparameters and training settings were established according to the default values used in the pre-training of each model with the COCO dataset, defined in [43]. The models were trained using two servers of a high performance computing cluster, both with 128 GB of RAM, one equipped with a 48-core Intel® Xeon® Silver 4214 CPU (2.20GHz) and two 32 GB NVIDEA® V100 GPUs, and other comprised of a 128-core AMD EPYC 7502 CPU and a 32 GB NVIDEA® V100 GPU.

C. MODEL EVALUATION

1) PERFORMANCE METRICS

The evaluation of the studied models was performed considering the mean average precision - mAP - and average recall - AR - metrics. These metrics reflect the models' accuracy and sensitivity for all object classes (on average) and for specific bounding box properties, such as the IOU threshold applied or the maximum number of detections. Further details concerning the computation of these metrics are provided in Appendix C.

During the training process, specific variations of these metrics were used for monitoring purposes, bearing in mind the particularities of the objects found in both datasets. Three specific quantities were analysed in detail to assess the algorithm's validation performance during training - $mAP@0.50IOU$, $AR@100$ or $AR@10$ (for the SIPAKMED and HFF regions studies, respectively), and total loss.

Although the values of mAP at larger IOU thresholds represent the algorithm's precision for detection with a more rigorous delineation of the object, the primary purpose of the models developed is to detect the cells or abnormal regions and not necessarily to delimit their exact boundaries. Thereafter, $mAP@0.50IOU$ allowed the assessment of the localisation performance of the network without being based on strict overlap criteria.

The analysis of the networks' detection sensitivity was supported by the average recall at a fixed maximum number of detections, which was selected according to the maximum number of objects per patch found in each dataset. The total loss was also analysed during training to monitor the state of the model's convergence.

2) MODEL'S CONVERGENCE EVALUATION

Despite the establishment of a fixed number of training steps, to anticipate the possible divergence of the model in the last training iterations, the network's validation performance was periodically evaluated (every 2000 steps) and the model with the best metric was the one considered final for the experiment in question.

Due to its representation of the overall robustness of the model, $mAP@0.50IOU$ was used as the comparison criterion to select the best training iteration for the HFF regions study. However, a distinct criterion - $AR@100$ was utilised in the case of the SIPAKMED dataset, because each of its images only contains ground truth annotations for cells of a single class, even though the captured microscopic field may include other cell types, which could lead to the incorrect assessment of some detected objects as false positives, considered in the computation of the mAP metric.

3) HYPERPARAMETER OPTIMISATION AND MODEL SELECTION

The validation performances of the best model obtained for the several cross-validation splits were averaged to derive the evaluation results of each experiment associated with a specific combination of hyperparameter values. The average cross-validation metrics achieved with the multiple hyperparameter combinations tested were compared to identify the best hyperparameter set-up for each model.

In the SIPAKMED study, the results attained with the optimal hyperparameter settings found for each detection model were compared to determine the most promising meta-architecture/backbone combination. Only the detection model with the best performance in the SIPAKMED dataset was contemplated in the experiments conducted in the HFF regions data.

Furthermore, to investigate the transferability of the knowledge acquired by the networks in the conventional microscopy domain to images of liquid-based preparations, two types of models (with the selected meta-architecture and backbone network) were explored: a **baseline model**, pre-trained on the COCO data; and a model **pre-trained with information from a conventional cytology domain** (the SIPAKMED dataset). These models were both re-trained in the mobile HFF regions dataset; a new hyperparameter search was performed, considering the same possible combinations for the two model types, to ensure the comparability of the experiments.

V. RESULTS AND DISCUSSION

A. SIPAKMED STUDY

This section contains the results referring to the models applied to the conventional cytology images of the SIPAKMED dataset to identify several single-cell types. The hyperparameter optimisation for each meta-architecture/backbone combination is firstly presented, providing a comparative analysis of the different models to identify the best one for this application domain. Then, the best-performing model is evaluated on the test set for the task of localisation and classification of cell types.

1) HYPERPARAMETER TUNING AND MODEL SELECTION

The average cross-validation detection performance obtained for the three models tested is supplied in Figure 5, along with the standard deviation of each metric. The average training time needed to attain the convergence of the models was approximately $8 - 26$ hours for SSD, $3 - 32$ hours for RetinaNet and $8 - 31$ hours for Faster R-CNN, being these times greatly influenced by the learning rate and patch size values selected for each experiment. The most advantageous hyperparameter combination for each model was determined based on the recall metric, for the reasons outlined in the previous section. As a consequence, the set-up of experiments 1, 5 and 3 was chosen as optimal for the SSD, RetinaNet and Faster R-CNN models, respectively.

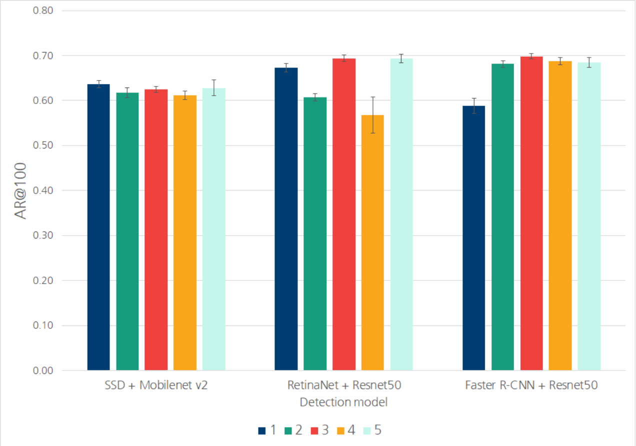


FIGURE 5. **Average 5-fold cross-validation *AR*@100 results obtained in the SIPAKMED dataset with the different hyperparameter settings tested for each meta-architecture/backbone combination. The represented values were obtained as the average validation results for the five cross-validation splits. The error bars correspond to the standard deviation over all splits.**

The networks trained with the optimal experimental setup for each meta-architecture/backbone combination were compared considering their detection ability for the distinct classes, evaluated separately using only the images annotated for the analysed class to account for the SIPAKMED dataset's peculiarities highlighted in section IV-C2 and illustrated in Figure 6. The $mAP@.50IOU$ and $AR@100$ values obtained are displayed in Figure 7.

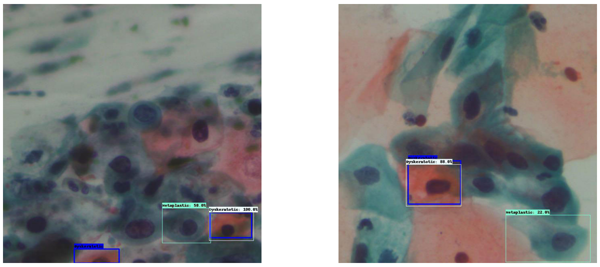
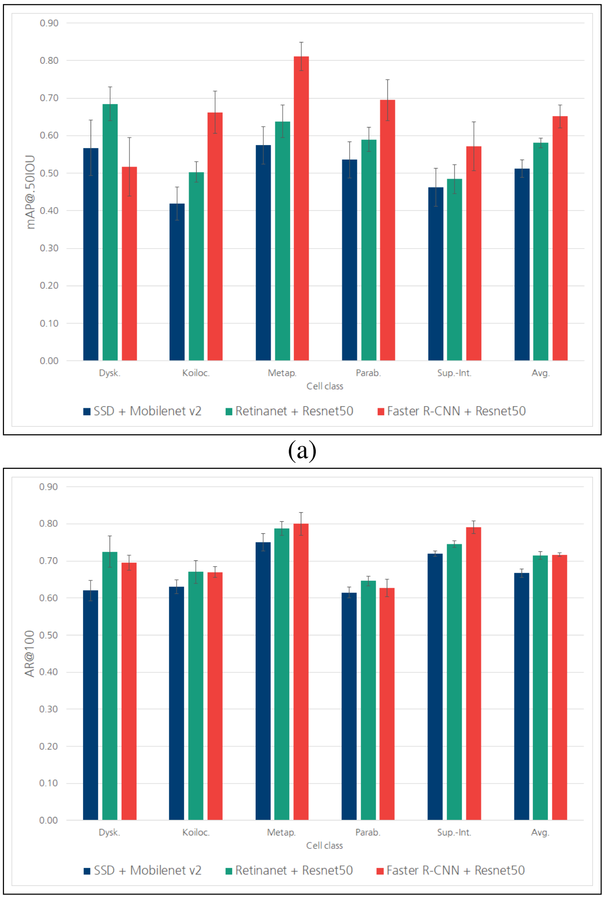


FIGURE 6. Examples of validation images for which the Faster R-CNN model detected cells of multiple types, despite having ground-truth information of a single cell type per image. The ground truth annotations are marked in dark blue, while light blue and white boxes outline the objects detected by the network. Correct detections can be identified when the dark blue and the lighter bounding boxes overlap.



(a)

(b)

FIGURE 7. Class-wise performance in the SIPAKMED dataset of the best hyperparameter set-up for each model, in terms of ***mAP*@.50*IOU*** (a) and ***AR*@100** (b). The represented values were obtained as the average validation results for the five cross-validation splits. The error bars correspond to the standard deviation over all splits.

The SSD model produced inferior outcomes in relation to the RetinaNet and Faster R-CNN detectors for both metrics, possibly a consequence of its simpler detection scheme and the lightweight backbone used (Mobilenet v2). Comparing the RetinaNet and Faster R-CNN models, both achieve similar average performances in terms of recall, translated in an equivalent sensitivity to the different cell types. Even though RetinaNet exhibits a better detection ability for dyskeratotic cells, the $mAP@.50IOU$ results indicate that Faster R-CNN was generally more robust. Thus, the Faster R-CNN architecture was identified as the most promising model, being the one used in the remaining experiments conducted for the mobile HFF regions dataset.

2) PERFORMANCE OF THE BEST MODEL

Table 3 comprehends the test results of the final Faster R-CNN model for the five classes, after being re-trained on the whole training set with the optimal hyperparameter set-up. Some examples of correct detections are presented in Figure 8.

TABLE 3. Class-wise performance in the SIPAKMED dataset of the best model, re-trained in the overall training set in terms of ***mAP*@.50*IOU*** and ***AR*@100**. The represented values were computed using the test data.

Class	Sup.-Int.	Parab.	Metap.	Dysk.	Koiloc.	Avg.
mAP@.50IOU	0.43956	0.35424	0.41867	0.34811	0.32930	0.37798
AR@100	0.68084	0.57389	0.71548	0.61549	0.59688	0.63651

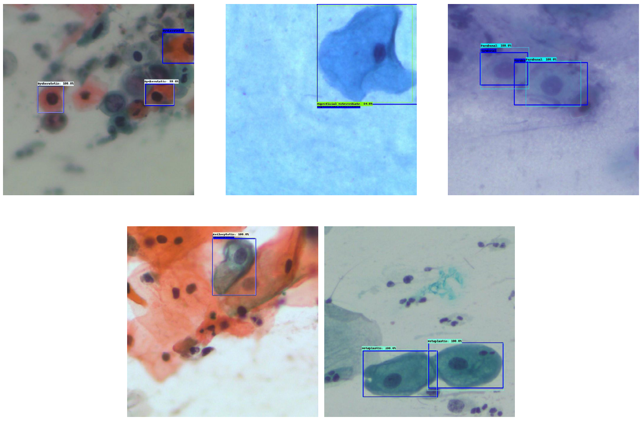


FIGURE 8. Examples of test images with cells correctly detected and classified by the Faster R-CNN model (true positives). The ground truth annotations are marked in dark blue, while light blue, green and white boxes outline the objects detected by the network.

On average, this model attained $mAP@.50IOU$ and $AR@100$ values of 0.37798 and 0.63651, correspondingly, demonstrating a performance deterioration from the validation data (which yielded $mAP@.50IOU$ and $AR@100$ values of 0.65120 and 0.71686) to the test set. The decline observed in the results might be due not only to a slight overfit to the training data, but also to the large amount of partial ground truth annotations (without actual cellular structures) that is more prevalent in the test data. These instances are a consequence of the division of the image field in patches, which leads to the division of annotated objects between adjacent patches and may in some cases result in ground truth objects that are not true objects of interest, as shown in

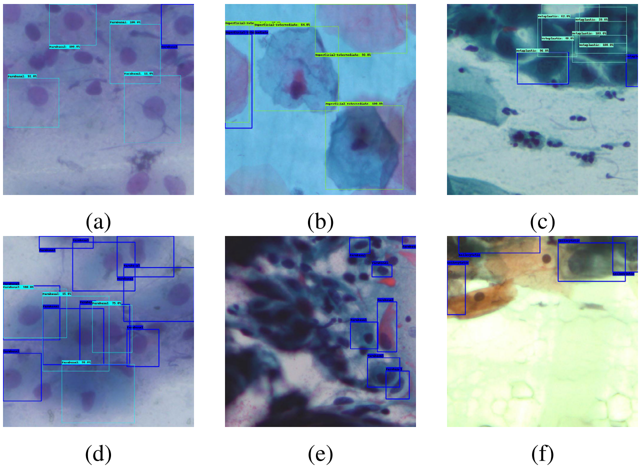

(a) (b) (c)
(d) (e) (f)

FIGURE 9. Examples of test images that highlight the performance limitations of the Faster R-CNN model. These comprise cells incorrectly detected by the network (false positives, in Figures 9(a) to 9(c)) and cells of interest that it missed (false negatives, in Figures 9(d) to 9(f)). The ground truth annotations are marked in dark blue, while light blue and green boxes outline the objects detected by the network.

Figure 9(f). Despite this decline, the class-wise metrics indicate that the model had superior detection ability for metaplastic and superficial-intermediate cells, which is coherent with the cross-validation results, highlighting the consistent behaviour of the network for different sets of data.

The confusion matrix of the detections obtained with this model is presented in Table 4. Regarding the discriminatory ability of the model, it was satisfactory in general (as attested by the correct detections for the five cell types in Figure 8), but some instances of dyskeratotic cells were misidentified as koilocytotic cells, potentially owing to the vesicular nuclei that characterise both cell types. In addition, a few koilocytotic cells were incorrectly classified as metaplastic, which might also have been a result of their similar properties, since these are two cell types found in the same region of the epithelium - the transformation zone - displaying enlarged nuclei and a transition morphology.

TABLE 4. Confusion matrix obtained with the final Faster R-CNN + Resnet50 model (re-trained in the overall training data with the hyperparameter set-up of experiment 3), evaluated for the SIPAKMED test set considering a score threshold of 0.1. Only detected objects were included in this analysis, disregarding false negatives.

		Predicted				
	Class	**Sup.-Int.**	**Parab.**	**Metap.**	**Dysk.**	**Koiloc.**
Ground truth	**Sup.-Int.**	27.21%	0.00%	0.00%	0.00%	0.00%
	Parab.	0.00%	18.71%	0.00%	0.17%	0.00%
	Metap.	0.34%	0.51%	14.12%	0.00%	0.68%
	Dysk.	0.00%	0.00%	0.17%	17.18%	2.55%
	Koiloc.	0.34%	0.00%	1.02%	0.34%	16.67%

Notwithstanding the aforementioned wrongly classified instances, in Table 4 it is possible to verify that these cases are a minor fraction of the network's predictions, revealing that the lack of precision evidenced by the lower test $mAP@.50IOU$ was mainly due to the false-positive detections, illustrated in Figures 9(a) to 9(c). The sub-optimal sensitivity of the model, materialised in the missed cells presented in Figures 9(d) to 9(f), was more accentuated in images whose properties hindered the individualisation of cell objects, such as the presence of cell aggregates (Figure 9(d)), dark fields of view (Figure 9(e)) or the partial obfuscation of the cellular structures (Figure 9(f)). The existence of ground-truth annotations without cellular structures also decreased the recall metric, even though it does not reflect a decline in the actual ability of the network to detect cells of interest.

In spite of the public availability of the SIPAKMED dataset and its application to segmentation and single-cell classification tasks, no approaches that used it to develop object detection models were found in the literature. Thus, the results reported in this article can be used as a performance benchmark for object detection methodologies, establishing a baseline for the development of future research in this field and corresponding to one of the main contributions of this work.

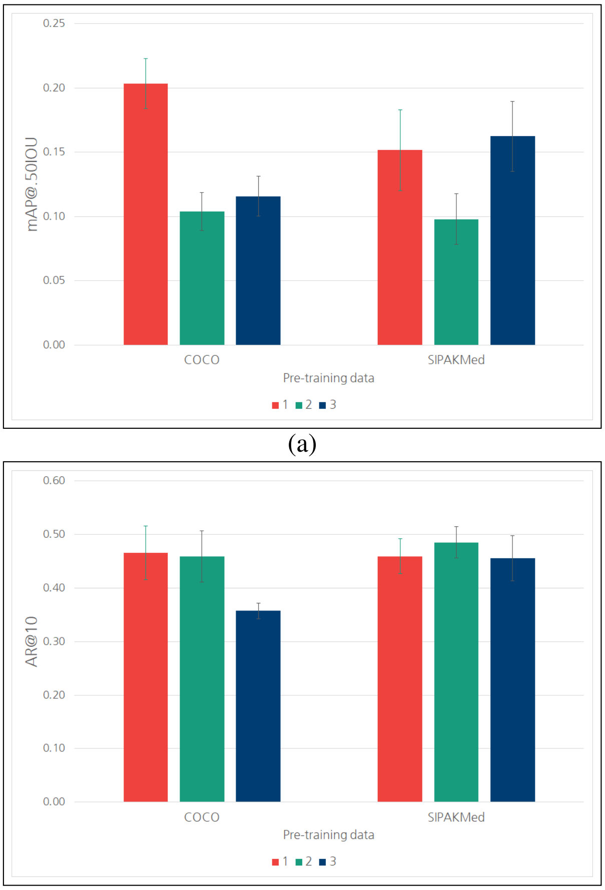
B. HFF REGIONS STUDY

This section is focused on the development of the model used to identify possible cervical lesions in LBC images, with the end goal of integrating it in a mobile IoT framework to support cervical lesion screening. The identification of the most adequate training settings is described in the first place, followed by a detailed analysis of the selected model.

1) HYPERPARAMETER TUNING AND MODEL SELECTION

The average cross-validation results obtained for the models pre-trained in either the COCO or the SIPAKMED datasets, and fine-tuned with HFF data, are provided in Figure 10. These models needed approximately $5 - 27$ hours to achieve convergence, depending on the hyperparameter values of the experiment. For the baseline model, pre-trained on the COCO dataset, the most adequate hyperparameter combination was the one tested in experiment 1, which culminated in a more precise model with an average cross-validation $mAP@.50IOU$ of 0.20315. Regarding the networks that took advantage of the SIPAKMED pre-training, all experimental set-ups allowed the training of models with equivalent detection sensitivities (visible in the $AR@10$ metric); still, the hyperparameter settings of experiment 3 yielded a superior precision, being selected as the final training set-up for this model.

The per-class performance of the model trained with the best hyperparameter setup for the two types of pre-training is detailed in Figure 11. Although the $AR@10$ values attained indicate that both models were identically sensitive to objects of interest, the detections of the network pre-trained in the COCO data were more accurate for all classes except ASC-H, which motivated the selection of this as the final model.



(a)

(b)

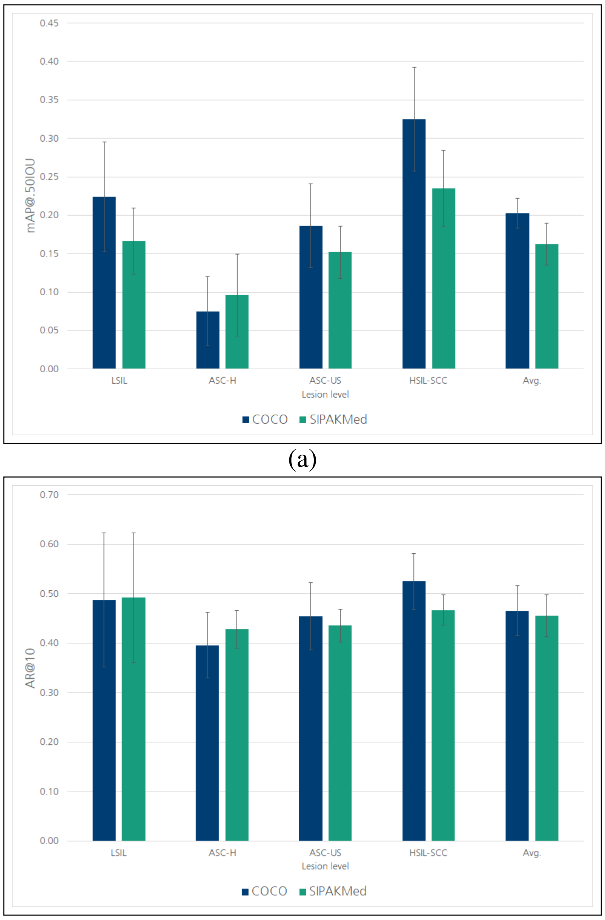
FIGURE 10. Average 5-fold cross-validation results achieved in the HFF regions dataset with the different hyperparameter settings tested for each type of pre-training, in terms of *mAP*@.50*IOU* (a) and *AR*@10 (b). The represented values were obtained as the average validation results for the five cross-validation splits. The error bars correspond to the standard deviation over all splits.

Therefore, the pre-training in the SIPAKMED data did not produce identifiable performance gains, evidencing that the knowledge transfer from a closer application domain was not advantageous in this case, contrary to what was expected. Whilst this can be the result of the circumscribed hyperparameter optimisation carried out or of the superior quality and size of the baseline COCO dataset, the substandard performances of both models suggest that the limitations of the HFF regions data had a sizeable impact on the networks' training.

2) FINAL MODEL ANALYSIS

An in-depth analysis was conducted for the Faster R-CNN model pre-trained on the COCO data and re-trained on the overall training set of the mobile HFF regions dataset. Table 5 contains the confusion matrix obtained for the test data, allowing the inspection of the model's classification ability.

The best classification performance was observed for ASC-US instances, the most represented class; however, some ASC-US cases were mistaken for other lesion levels, possibly a product of the different cellular changes encompassed by this level. In contrast, the null values obtained for LSIL were a result of the lack of detections of this class that matched any ground truth annotations, revealing that its under-representation in the training data prevented the network from learning its typical features.



(a)

(b)

FIGURE 11. Class-wise performance in the mobile HFF regions dataset of the best hyperparameter set-up for both types of pre-training, in terms of *mAP*@.50*IOU* (a) and *AR*@10 (b). The represented values were obtained as the average validation results for the five cross-validation splits. The error bars correspond to the standard deviation over all splits.

TABLE 5. Confusion matrix obtained with the final Faster R-CNN + Resnet50 model (re-trained in the overall training data with the hyperparameter set-up of experiment 1), evaluated for the HFF regions test set considering a score threshold of 0.1. Only detected objects were included in this analysis, disregarding false negatives.

		Predicted			
	Class	**ASC-US**	**LSIL**	**ASC-H**	**HSIL-SCC**
Ground truth	**ASC-US**	41.67%	8.33%	8.33%	2.78%
	LSIL	0.00%	0.00%	0.00%	0.00%
	ASC-H	2.78%	0.00%	11.11%	0.00%
	HSIL-SCC	2.78%	2.78%	13.89%	5.56%

Nonetheless, the model presented a reasonable capability to discriminate regions of the ASC-H and HSIL-SCC classes, and most of the incorrectly classified instances were misidentified as belonging to adjacent lesion levels, which may exhibit common properties that are a consequence of the progressive nature of cervical lesions. For these reasons, it is possible to say that the model demonstrated a satisfactory classification performance.

More information regarding the model's robustness for each class is provided in Table 6. The reported metrics were substandard overall, with the lowest performance being observed for the LSIL class, whose representation in the dataset was insufficient. The lesion level more accurately detected by the network was HSIL-SCC, due to its more distinctive features, in association with the larger dimensions of the regions of interest and higher amount of training data available.

TABLE 6. Class-wise performance of the Faster R-CNN + Resnet50 model (pre-trained on the COCO dataset and re-trained in the overall HFF regions training data with the hyperparameter set-up of experiment 1), in terms of *mAP*@.50*IOU* and *AR*@10, evaluated for the HFF regions test set.

Class	LSIL	ASC-H	ASC-US	HSIL-SCC	Avg.
mAP@.50IOU	0.000744	0.00706	0.008573	0.03836	0.01368
AR@10	0.13125	0.27692	0.25754	0.32973	0.24886

In comparison with the cross-validation results, this model achieved a poorer performance on the test data, obtaining $mAP@.50$ and $AR@10$ values of just 0.01368 and 0.24886. Although the network was able to identify correctly some of the lesion instances, exemplified in Figure 12, there is a discrepancy of its robustness for the validation and test data that might be a consequence of the applied pre-processing pipeline and the dataset's properties. This corroborates what was found in [26], which investigated the impact of the training data volume in the test performance of detection networks and demonstrated that the reduction of the dataset to a subset of images (containing approximately 4500 annotated regions out of 33500) substantially impaired the robustness of their baseline detection model. Their baseline model - also a Faster R-CNN with a Resnet50 backbone - trained on the smaller training set yielded test metrics in the same order of magnitude (mAP and AR values of 0.066 and 0.129) as the ones obtained in this work for the HFF regions dataset. Even though a direct comparison is not possible, due to the different variations of the mAP and AR metrics considered and the distinct test datasets used, the achievement of similar values in spite of the lower amount of training instances available (927) confirms that the proposed pipeline has encouraging potential for the development of a cervical cancer screening system.

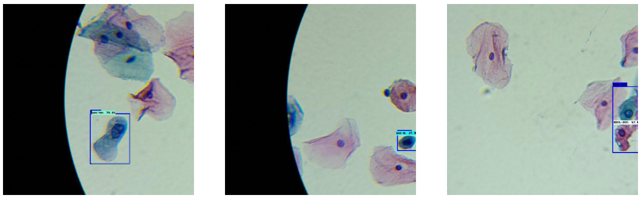


FIGURE 12. Examples of test images with cervical lesion regions correctly detected and classified by the Faster R-CNN model (true positives) in the mobile HFF regions dataset. The ground truth annotations are marked in dark blue, while light blue, green and white boxes outline the objects detected by the network.

Similarly to what was verified for the SIPAKMED data, a possible cause for the lower test metrics is the existence of ground truth annotations without any cellular structures, much more numerous in the test set, which are accounted as false negatives (Figure 13(g)), decreasing the evaluated $AR@10$ performance of the network with no true impact in its detection sensitivity. Another factor that may have contributed to the performance deterioration is the diversity of morphological properties that is associated with each cervical lesion level, which was not properly learned by the network, owing to the scarce volume of data available and the imbalance of the different classes of lesions. In addition, many of the model's false positive or wrongly classified detections, e.g. Figures 13(a) to 13(c),, exhibit properties that are in fact characteristic of the predicted lesion levels; these may correspond to dubious cases that were not annotated for that reason but are still clinically relevant. Finally, the presence of other types of structures that were not included in the annotation classes but have distinctive features in common with some cervical lesion types, such as inflammatory cells (Figure 13(h)), also increases the false-positive detections of the model, decreasing its precision.

In spite of the variability of the structures in the dataset and the aforementioned limitations, the network was still able to identify patterns in the data for each lesion class. Hence, although it was not manifested in the mAP and AR values, the detection ability of the model demonstrated a good potential for integration in a decision support system to assist medical professionals, which could help them in the recognition of lesion candidates from microscopic fields and subsequently improve their diagnosis sensitivity, as studied in [44].

To further assess the applicability of the proposed approach in a practical setting, the execution time of the main steps of the pipeline was evaluated separately and combined to approximate the time needed to analyse an entire cervical sample. The pipeline proposed requires approximately 4 minutes to process an entire sample on an equipment with good computational capabilities (high performance computing server comprising 128GB of RAM, one NVIDEA V100 GPU with 16GB and an Intel(R) Xeon(R) Silver 4114 CPU with 40 cores @ 2.20GHz), being most of this time consumed by the pre-processing steps: the segmentation of the optical circle and the division of each image into several patches required on average $2.768s$ per image, while the detection of regions in all the patches of each image only demanded $0.50773s$. This indicates a promising potential for deployment situations in which network connection to remote

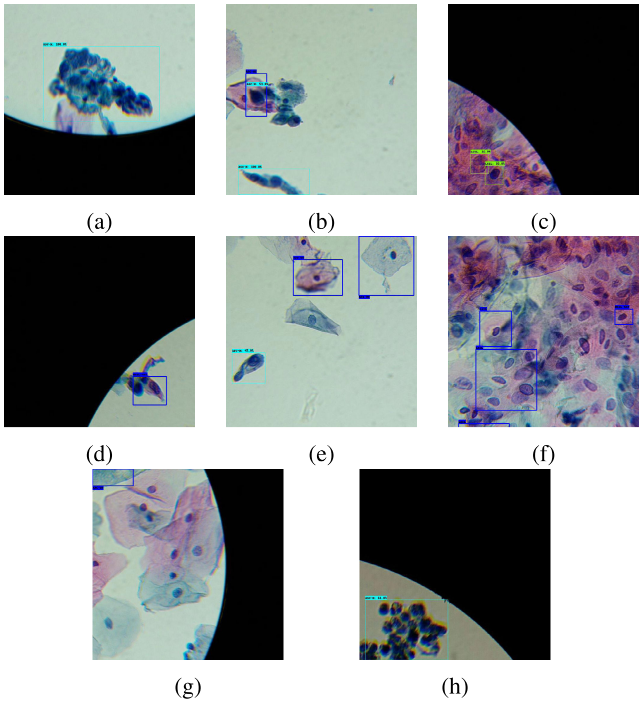


FIGURE 13. Examples of incorrect detections predicted by the Faster R-CNN model for the HFF regions test data. These comprise regions with properties coherent with the predicted lesion levels that did not match any ground truth annotation (considered false positives, in Figures 13(a) to 13(c)), cells of interest that it missed (false negatives, in Figures 13(d) to 13(f)) and other incorrect detections that highlight the limitations of the developed pipeline (Figures 13(g) and 13(h). The ground truth annotations are marked in dark blue, while light blue and green boxes outline the objects detected by the network.

servers is available. Nonetheless, this estimation does not take into consideration the image acquisition time, which is approximately 30-40 minutes for the current version of the μSmartScope device, corresponding to the main bottleneck of the pipeline. In the future, the pipeline execution time in mobile devices should also be investigated, to explore the possibility of mobile deployment without the need for external computational resources.

VI. CONCLUSION AND FUTURE WORK

This paper proposed the integration of deep object detection networks in an IoT-based system to analyse microscopic fields of liquid-based cytology samples and locate regions indicative of cervical lesions.

The initial study performed on the public SIPAKMED database enabled the validation of the detection framework, as the optimised model obtained a satisfactory performance in the location and recognition of five cervical cell types from conventional cytology images. Still, a more adequate patch extraction pipeline could contribute to the development of more accurate models and should be considered in future refinements of the overall pipeline. Nevertheless, to the knowledge of the authors, this was the first study that reported results for the cell detection task in the SIPAKMED data, providing a performance benchmark for future research in this field.

For the cervical lesion detection task, transfer learning from the SIPAKMED model was explored as a way of addressing the shortcomings of the liquid-based cytology dataset acquired with the mobile-based microscope. However, it did not provide a performance advantage to the trained networks, in comparison with another public database from an unrelated application domain, possibly due to a sub-optimal hyperparameter optimisation, to the superior quality of the COCO dataset used in the baseline experiments, or to the limitations of the LBC data. Notwithstanding, the knowledge transfer from the conventional cytology domain corresponds to an innovative strategy that can be refined (through the optimisation of the fine-tuned layers, for instance) to enable appreciable performance gains.

The final model developed for the analysis of LBC images yielded sub-standard test metrics, indicating that its training and evaluation were affected by the subjectivity of the cervical lesion assessment procedure and by the shortcomings of the mobile HFF regions dataset. Therefore, to achieve a model sufficiently robust for integration in a clinical decision support system, future studies should consider these concerns. A promising approach can be the exploration of hybrid pipelines with inter-dependent modules responsible for separating the heterogeneous cell types that characterise similar lesion classes. Concurrently, the dataset must be extensively improved in terms of data volume and class representation. Moreover, the subjectivity of the annotation process should be addressed, directing efforts towards the standardisation of the identified objects of interest, either through the provision of analysis guidelines, the execution of inter-expert agreement studies or the stratification of the TBS lesion levels into more classes that enable a better recognition of typical features.

Although further improvements are required to ensure an adequate reliability for deployment in a clinical context, the explored approach exhibits promising results (comparable to the ones attained in similarly data-limited conditions) that highlight its potential for integration in a diagnosis workflow with an expert on the loop, as a way of assisting the medical decision. Even though a suitable interface to provide feedback to the users still needs to be implemented, the studies here conducted provide useful recommendations to guide future research in this area, which, together with the completely automated image acquisition module based on the μSmartScope device, correspond to important steps forward in the development of a cost-effective mobile IoT framework that supports cervical lesion screening.

DATA AVAILABILITY STATEMENT

Due to concerns regarding the privacy of the involved patients, the cervical lesions dataset used in the conducted studies shall remain private. Nevertheless, efforts are currently being carried out by the research team to provide a public dataset based on similar data.

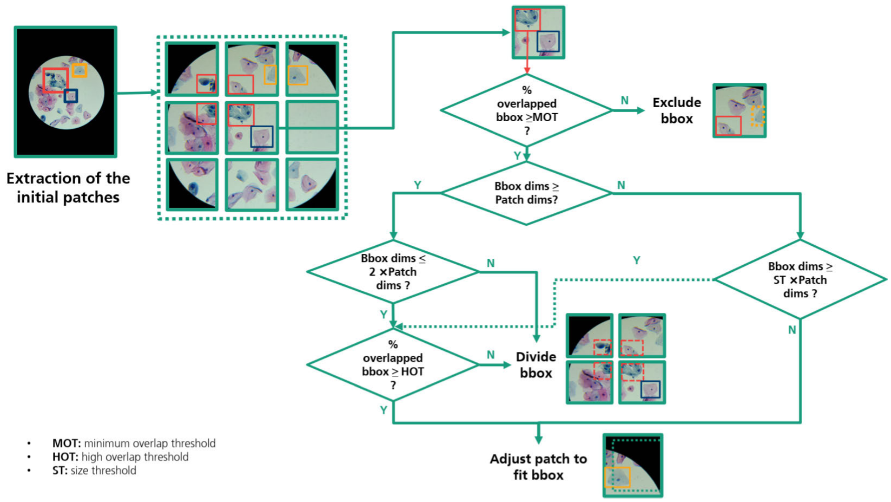


FIGURE 14. Methodology used to adjust the extracted patches to the associated bounding boxes.

APPENDIX A
DETAILED DESCRIPTION OF THE PATCH EXTRACTION PROCESS

The procedure followed to extract the image patches took into account the properties and limitations of the dataset, in particular the fact that the sample fields are mainly concentrated in the centre of the image and the presence of bounding box annotations, which must be considered when dividing the image into several regions. The extraction of patches of pre-established dimensions comprised several steps:

1) Extraction of the initial patches as the ones that yielded the maximum amount of adjacent patches within the image's boundaries, centred in the optic disk region. If the image dimensions were not entirely divisible by the selected patch size, a part of the source image's dark frame area, equally distributed in all directions, was also considered as the patch extraction region, to ensure that no regions from the actual microscopic field were left out of the analysis by the model.

2) Adjustment of the region encompassed by each patch according to the annotations that overlapped with it. As each bounding box could be present in several patches, some bounding box inclusion and patch adjustment criteria were taken into account, being described in detail in the flowchart of Figure 14. However, this step was only applied to the training data; the test patches were not adjusted to the bounding boxes, to mimic the inference process that will be performed by the model in new data and subsequently allow an independent performance assessment in the test set.

3) Resizing the final patch to fit the desired output dimensions, to guarantee that all the images provided to the network had the same size.

In the end, each subset was assembled considering all the patches from the images that were attributed to that specific subset.

APPENDIX B
CHARACTERISATION OF THE PATCH DOWN-SAMPLING STEPS

As the hardware constraints only allowed the usage of small mini-batch sizes to train the models, the predominance of empty patches in the training set would lead to an ineffective learning process: in many iterations the weight updates would be performed using only information from empty images, with very few information regarding the localization and classification of the objects, hindering a correct convergence of the network. Subsequently, reducing the proportion of empty images considered in each training iteration was pivotal for a successful optimisation of the models.

This process was controlled by two parameters: the ratio between empty and annotated patches and the ratio of empty patches in relation to the total amount of patches. For each original image, if there were any annotated patches, the number of considered empty patches was defined according to the first ratio; in case all patches of that image were empty, the number of patches added to the final training set was determined as a percentage of the total amount of patches extracted from that image (second ratio). Considering the number of empty patches to keep in the training set, the patch instances actually added were randomly sampled from the empty instances of that image. A minimum of 1 empty patch per source image was added, to guarantee sufficient diversity of the empty microscopic fields. This process was only applied to the training sets, so that the evaluation was as similar as possible to the usage of the model in the deployment scenario. The percentage of annotated patches was maximised for an empty to annotated ratio of 0.25 for the two patch dimensions considered, with an optimal empty to total ratio of 0.05 for the 320 × 320 patches and an empty to total ratio of 0.15 for the 640 × 640 patches, since this parameter did not influence the final amount of empty patches used.

APPENDIX C
COMPUTATION OF THE EVALUATION METRICS

Given that object detection involves two separate sub-tasks - the localisation of the objects and the determination of their class -, it is important to evaluate the developed algorithms according to metrics that reflect their performance in both tasks. Therefore, the detection networks were evaluated using the main metrics reported for the PASCAL Visual Object Classes (VOC) [29] and COCO challenges [30], namely mean average precision and recall.

Most metrics used to evaluate detection algorithms are based on the intersection over union (IoU) of the predicted bounding boxes in relation to the ground truth boxes, computed as the ratio between the intersection of both boxes and their union [45]. The IoU between two arbitrary regions A and B can be calculated according to eq. (1). The determination of the algorithm's true positive (TP), false positive (FP) and false negative (FN) detections is achieved by establishing a fixed IoU threshold, enabling the estimation of common machine learning metrics, such as precision and recall [46] (represented in eqs. (2) and (3)).

$$IoU = \frac{A \cap B}{A \cup B} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Average precision (AP), considered in the Pascal VOC challenge, is computed from the model's precision-recall curve as the area under the monotonically decreasing version of the curve. In this case, the precision-recall curve is obtained using a fixed IoU threshold of 0.5, by varying the threshold of the detections' confidence value. In contrast, to determine COCO's mean average precision (mAP), several AP values are computed by averaging the precision values for all the recall levels found during the evaluation process. Each AP is associated with a distinct IoU threshold in the $0.5-0.95$ range (using a step of 0.05), and the mean of the resulting AP values for all the object categories is computed to attain the mAP score.

Average recall (AR), also considered in the COCO challenge, corresponds to the mean of the recall values of all the object classes resulting from IoU thresholds in the $0.5 - 1.0$ range, for a fixed number of detections [47]. AR can be determined for distinct numbers of detections and object scales [30]. Both AP and AR are computed firstly for each class and then averaged over all classes to yield the corresponding values.

APPENDIX D
HYPERPARAMETER SET-UPS CONSIDERED IN THE PERFORMED EXPERIMENTS

A. SIPAKMED EXPERIMENTS

See Table 7.

TABLE 7. Hyperparameter values tested for the five experiments conducted for the meta-architecture/backbone combinations in the SIPAKMED study.

SSD model with a Mobilenet v2 backbone			
Experiment ID	Batch size	Learning rate	Image size
1	8	2.00923E-04	320
2	16	1.09750E-06	320
3	16	2.78256E-06	640
4	8	2.25702E-03	640
5	8	1.96304E-05	320
RetinaNet model with a Resnet50 backbone			
Experiment ID	Batch size	Learning rate	Image size
1	16	1.45083E-06	320
2	16	1.91791E-06	640
3	8	1.26186E-04	320
4	8	4.64159E-04	640
5	16	2.00923E-04	320
Faster R-CNN model with a Resnet50 backbone			
Experiment ID	Batch size	Learning rate	Image size
1	16	7.56463E-03	640
2	8	4.97702E-05	320
3	8	1.07227E-03	640
4	16	1.02353E-05	320
5	16	1.55568E-03	320

B. HFF REGIONS EXPERIMENTS

See Table 8.

TABLE 8. Hyperparameter values tested for the three experiments conducted for the Faster R-CNN model with a Resnet50 backbone in the HFF regions study.

Experiment ID	Batch size	Learning rate	Image size
1	8	1.38489E-04	640
2	8	3.35160E-06	320
3	16	6.89261E-03	640

ACKNOWLEDGMENT

The authors would like to thank the Anatomical Pathology Service of Hospital Prof. Doutor Fernando Fonseca Pathology Anatomy Service.

REFERENCES

[1] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjosé, M. Saraiya, J. Ferlay, and F. Bray, "Estimates of incidence and mortality of cervical cancer in 2018: A worldwide analysis," *Lancet Global Health*, vol. 8, no. 2, pp. e191–e203, Feb. 2020, doi: 10.1016/S2214-109X(19)30482-6.

[2] *World Health Organization, and Reproductive Health and Research, Comprehensive Cervical Cancer Control: A Guide to Essential Practice*, World Health Org., Geneva, Switzerland, 2014.

[3] Y. Jusman, S. C. Ng, and N. A. Abu Osman, "Intelligent screening systems for cervical cancer," *Sci. World J.*, vol. 2014, pp. 1–15, Jan. 2014, Art. no. e810368, doi: 10.1155/2014/810368.

[4] T. Conceição, C. Braga, L. Rosado, and M. J. M. Vasconcelos, "A review of computational methods for cervical cells segmentation and abnormality classification," *Int. J. Mol. Sci.*, vol. 20, no. 20, p. 5114, Oct. 2019, doi: 10.3390/ijms20205114.

[5] M. J. M. Vasconcelos, D. Elias, J. M. C. da Costa, and J. S. Cardoso, "μSmartScope: Towards a fully automated 3D-printed smartphone microscope with motorized stage," in *Biomedical Engineering Systems and Technologies*. Cham, Switzerland: Springer, 2017, pp. 19–44, doi: 10.1007/978-3-319-94806-5_2.

[6] L. Rosado, J. M. C. D. Costa, D. Elias, and J. S. Cardoso, "Automated detection of malaria parasites on thick blood smears via mobile devices," *Proc. Comput. Sci.*, vol. 90, pp. 138–144, Jan. 2016, doi: 10.1016/j.procs.2016.07.024.

[7] L. Rosado, J. da Costa, D. Elias, and J. Cardoso, ''Mobile-based analysis of malaria-infected thin blood smears: Automated species and life cycle stage determination,'' *Sensors*, vol. 17, no. 10, p. 2167, Sep. 2017, doi: 10.3390/s17102167.

[8] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, ''Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images,'' in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3144–3148, doi: 10.1109/ICIP.2018.8451588.

[9] M. M. Rahaman, C. Li, X. Wu, Y. Yao, Z. Hu, T. Jiang, X. Li, and S. Qi, ''A survey for cervical cytopathology image analysis using deep learning,'' *IEEE Access*, vol. 8, pp. 61687–61710, 2020, doi: 10.1109/ACCESS.2020.2983186.

[10] N. B. Byju, V. K. Sujathan, P. Malm, and R. R. Kumar, ''A fast and reliable approach to cell nuclei segmentation in PAP stained cervical smears,'' *CSI Trans. ICT*, vol. 1, no. 4, pp. 309–315, Dec. 2013, doi: 10.1007/s40012-013-0028-y.

[11] Z. Lu, G. Carneiro, A. P. Bradley, D. Ushizima, M. S. Nosrati, A. G. Bianchi, C. M. Carneiro, and G. Hamarneh, ''Evaluation of three algorithms for the segmentation of overlapping cervical cells,'' *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 441–450, Mar. 2016, doi: 10.1109/JBHI.2016.2519686.

[12] W. William, A. Ware, A. H. Basaza-Ejiri, and J. Obungoloch, ''A pap-smear analysis tool (PAT) for detection of cervical cancer from pap-smear images,'' *Biomed. Eng. OnLine*, vol. 18, no. 1, pp. 1–22, Dec. 2019, doi: 10.1186/s12938-019-0634-5.

[13] S. Gautam, A. Bhavsar, A. K. Sao, and K. Harinarayan, ''CNN based segmentation of nuclei in PAP-smear images with selective preprocessing,'' *Proc. SPIE*, vol. 10581, Mar. 2018, Art. no. 105810X, doi: 10.1117/12.2293526.

[14] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, ''Pap-smear benchmark data for pattern classification,'' in *Nature inspired Smart Information Systems*. Albufeira, Portugal: NiSIS, 2005, pp. 1–9. [Online]. Available: https://orbit.dtu.dk/en/publications/pap-smear-benchmark-data-for-pattern-classification

[15] N. C. F. B. Information, U. S. N. L. O. M. R. Pike, B. MD, and. Usa. *Cancer and Pre-Cancer Classification Systems*. World Health Organization, Publication Title: Comprehensive Cervical Cancer Control: A Guide to Essential Practice. 2nd Edition. Accessed: Aug. 2, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK269605/

[16] S. Fekri-Ershad, ''Pap smear classification using combination of global significant value, texture statistical features and time series features,'' *Multimedia Tools Appl.*, vol. 78, no. 22, pp. 31121–31136, Nov. 2019, doi: 10.1007/s11042-019-07937-y.

[17] J. V. Lorenzo-Ginori, W. Curbelo-Jardines, J. D. López-Cabrera, and S. B. Huergo-Suárez, ''Cervical cell classification using features related to morphometry and texture of nuclei,'' in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (Lecture Notes in Computer Science), J. Ruiz-Shulcloper and G. S. di Baja, Eds. Berlin, Germany: Springer, 2013, pp. 222–229.

[18] A. Arora, A. Tripathi, and A. Bhan, ''Classification of cervical cancer detection using machine learning algorithms,'' in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 827–835, doi: 10.1109/ICICT50816.2021.9358570.

[19] A.-U. Rehman, N. Ali, I. Taj, M. Sajid, and K. S. Karimov, ''An automatic mass screening system for cervical cancer detection based on convolutional neural network,'' *Math. Problems Eng.*, vol. 2020, Oct. 2020. [Online]. Available: https://www.hindawi.com/journals/mpe/2020/4864835/

[20] H. Wang, C. Jiang, K. Bao, and C. Xu, ''Recognition and clinical diagnosis of cervical cancer cells based on our improved lightweight deep network for pathological image,'' *J. Med. Syst.*, vol. 43, no. 9, pp. 1–9, Sep. 2019, doi: 10.1007/s10916-019-1426-y.

[21] M. Kwon, M. Kuko, M. Pourhomayoun, V. Martin, T. Kim, and S. Martin, ''Multi-label classification of single and clustered cervical cells using deep convolutional networks,'' in *Proc. Int. Conf. Data Sci. (ICDATA)*, 2018, pp. 10–15.

[22] R. Nayar and D. C. Wilbur, *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*. Cham, Switzerland: Springer, 2015, doi: 10.1007/978-3-319-11074-5.

[23] Du, X. Li, and Q. Li, ''Detection and classification of cervical exfoliated cells based on faster R-CNN,'' in *Proc. IEEE 11st Int. Conf. Adv. Infocomm Technol. (ICAIT)*, Oct. 2019, pp. 52–57.

[24] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[25] X. Tan, K. Li, J. Zhang, W. Wang, B. Wu, J. Wu, X. Li, and X. Huang, ''Automatic model for cervical cancer screening based on convolutional neural network: A retrospective, multicohort, multicenter study,'' *Cancer Cell Int.*, vol. 21, no. 1, pp. 1–10, Dec. 2021, doi: 10.1186/s12935-020-01742-6.

[26] Y. Liang, Z. Tang, M. Yan, J. Chen, Q. Liu, and Y. Xiang, ''Comparison-based convolutional neural networks for cervical Cell/Clumps detection in the limited data scenario,'' 2018, *arXiv:1810.05952*.

[27] X. Zhu, X. Li, K. Ong, W. Zhang, W. Li, L. Li, D. Young, Y. Su, B. Shang, L. Peng, and W. Xiong, ''Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears,'' *Nature Commun.*, vol. 12, no. 1, p. 3541, Dec. 2021, doi: 10.1038/s41467-021-23913-3.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, and L. Fei-Fei, ''ImageNet large scale visual recognition challenge,'' *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[29] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, ''The PASCAL visual object classes (VOC) challenge,'' *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, ''Microsoft COCO: Common objects in context,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48

[31] H. A. Phoulady and P. R. Mouton, ''A new cervical cytology dataset for nucleus detection and image classification (Cervix93) and methods for cervical nucleus detection,'' 2018, *arXiv:1811.09651*.

[32] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, ''Speed/accuracy trade-offs for modern convolutional object detectors,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7310–7311.

[33] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[34] X. Wu, D. Sahoo, and S. C. H. Hoi, ''Recent advances in deep learning for object detection,'' 2019, *arXiv:1908.03673*.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, ''MobileNetV2: Inverted residuals and linear bottlenecks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, ''SSD: Single shot multibox detector,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46448-0_2, doi: 1007/978-3-319-46448-0_2.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, ''Focal loss for dense object detection,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2018.

[38] J. Bergstra and Y. Bengio, ''Random search for hyper-parameter optimization,'' *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 1–25, 2012.

[39] Y. Bengio, ''Practical recommendations for gradient-based training of deep architectures,'' in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2012, pp. 437–478. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-35289-8_26

[40] A. F. Sampaio, J. Gonçalves, L. Rosado, and M. J. M. Vasconcelos. *Cluster-Based Anchor Box Optimisation Method for Different Object Detection Architectures*. Accessed: Jul. 12, 2021. [Online]. Available: https://recpad2020.uevora.pt/wp-content/uploads/2020/10/RECPAD_2020_paper_42.pdf

[41] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[42] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[43] (Jul. 12, 2021). *Configuration Settings for the Models of the Tensorflow Object Detection API*. [Online]. Available: https://github.com/tensorflow/models/tree/c9c230d97fa54c7165ad31d663033%ba83ce610e4/research/object_detection/samples/configs

[44] H. Bao, X. Sun, Y. Zhang, B. Pang, H. Li, L. Zhou, F. Wu, D. Cao, J. Wang, B. Turic, and L. Wang, "The artificial intelligence-assisted cytology diagnostic system in large-scale cervical cancer screening: A population-based cohort study of 0.7 million women," *Cancer Med.*, vol. 9, no. 18, pp. 6896–6906, Sep. 2020, doi: 10.1002/cam4.3296.

[45] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," 2019, *arXiv:1902.09630*.

[46] D. Powers, "Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation," School Inform. Eng., Flinders Univ., Adelaide, SA, Australia, Tech. Rep. SIE-07-001, Dec. 2007, p. 24.

[47] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 814–830, Apr. 2016, doi: 10.1109/TPAMI.2015.2465908.

ANA FILIPA SAMPAIO received the M.Sc. degree in biomedical engineering from the Faculty of Engineering, University of Porto, in 2019.

She has experience in the area of biomedical signal processing, acquired through an Erasmus+ internship at the University of Twente, and in image analysis and deep learning methodologies, in-depth skills developed in the context of her dissertation. Her knowledge regarding computer vision and machine learning allowed her to play the role of a Monitor of the Curricular Unit of Computer Vision, Faculty of Engineering, University of Porto. Since 2019, she has been a Researcher with the Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions (AICOS). Her main research interests include the application of computer vision and machine learning to healthcare, namely for the development of medical monitoring and decision support systems.

LUÍS ROSADO received the M.Sc. degree in biomedical engineering from the Instituto Superior Técnico, Technical University of Lisbon, in 2009, and the Ph.D. degree in biomedical engineering from the Faculty of Engineering, University of Porto, in 2018.

He has worked as a Researcher with Reverse Engineering, being involved in the development of computer vision tools for the analysis of biological experiments and inspection of critical infrastructures. Since 2011, he has been a Researcher with the Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions (AICOS), focusing its research activity in the areas of computer vision and machine learning, where he is currently a Senior Scientist. During this period, he has been applying those competencies in the development of solutions for healthcare and auditing, particularly in the development of computer-aided diagnosis (CADx) and quality control systems, respectively. His main research interests include the development of mobile solutions based on computer vision and artificial intelligence, where he actively works in the development of computer-aided methods for detection of different pathologies on microscopic images, development of automated mobile-based optical devices, and automated approaches for detection and counting of vector insects responsible for key pests in viticulture.

MARIA JOÃO M. VASCONCELOS received the B.Sc. degree in mathematics applied to technology from the Faculty of Sciences, University of Porto, in 2002, and the master's degree in applied statistics and modelling, and the Ph.D. degree in informatics engineering from the Faculty of Engineering, University of Porto, in 2006 and 2015, respectively.

She worked in statistics with Interpay Nederland BV and as a Consultant of document management with Link Consulting SA. She also worked as an Invited Assistant with the Faculty of Engineering, University of Porto, and the School of Engineering, Polytechnic of Porto. She is currently a Senior Scientist with the Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions (AICOS). Her main research interests include digital image processing and computer vision.